School of Computer Science and Engineering

(Computer Science & Engineering)

Faculty of Engineering & Technology

Jain Global Campus, Kanakapura Taluk - 562112
Ramanagara District, Karnataka, India

2023-2024

(VIII Semester)

A Project Report on

"120 Years of Grandeur: Unveiling Trends and Insights in Olympic data"

Submitted in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

Hitha Choudhary G, K Shreeshanth Gouda, Lakshya Sharma
22BTRAD015, 22BTRAD017, 22BTRAD021

Under the guidance of

Mr. Akash Das

AVP and Project Manager

Futurense Technologies

FACULTY OF ENGINEERING AND TECHNOLOGY

Department of Computer Science and Engineering

School of Computer Science & Engineering

Faculty of Engineering & Technology

Jain Global Campus, Kanakapura Taluk - 562112
Ramanagara District, Karnataka, India

CERTIFICATE

This is to certify that the project work titled **"120 Years of Grandeur: Unveiling Trends and Insights in Olympic data"** is carried out by **Hitha Choudhary G (22BTRAD015), K Shreeshanth Gouda (22BTRAD017), Lakshya Sharma (22BTRAD021),** a bonafide student(s) of Bachelor of Technology at the School of Engineering & Technology, Faculty of Engineering & Technology, JAIN (Deemed-to-be University), Bangalore in partial fulfillment for the award of degree in Bachelor of Technology in Computer Science and Engineering, during the year **2023-2024**.

Mr. Akash Das
AVP and Project Manager
Futurense Technologies

Date: 10-04-2024

Dr. Aditya Pai,
Program Head,
Computer Science and Engineering,
School of Computer Science & Engineering
Faculty of Engineering & Technology
JAIN (Deemed to-be University)

Date: 10-04-2024

Dr. Geetha G
Director,
School of Computer Science & Engineering
Faculty of Engineering & Technology
JAIN (Deemed to-be University)

Date: 10-04-2024

Name of the Examiner

Signature of Examiner

1.

2.

DECLARATION

We, **Hitha Choudhary G (22BTRAD015), K Shreeshanth Gouda (22BTRAD017), Lakshya Sharma (22BTRAD021)** student of IV semester B.Tech in **Computer Science and Engineering**, at School of Engineering & Technology, Faculty of Engineering & Technology, **JAIN (Deemed to-be University)**, hereby declare that the internship work titled **"120 Years of Grandeur: Unveiling Trends and Insights in Olympic data"** has been carried out by us and submitted in partial fulfilment for the award of degree in **Bachelor of Technology in Computer Science and Engineering** during the academic year **2023-2024**. Further, the matter presented in the work has not been submitted previously by anybody for the award of any degree or any diploma to any other University, to the best of our knowledge and faith.

Name1: Hitha Choudhary G Signature

USN : 22BTRAD015

Name2: K Shreeshanth Gouda Signature

USN : 22BTRAD017

Name3: Lakshya Sharma Signature

USN : 22BTRAD021

Place : Bangalore

Date : 10-04-2024

ACKNOWLEDGEMENT

It is a great pleasure for me to acknowledge the assistance and support of a large number of individuals who have been responsible for the successful completion of this project work.

First, I take this opportunity to express my sincere gratitude to Faculty of Engineering & Technology, JAIN (Deemed to-be University) for providing me with a great opportunity to pursue my Bachelors Degree in this institution.

I am deeply thankful to several individuals whose invaluable contributions have made this project a reality. I wish to extend my heartfelt gratitude to ***Dr. Chandraj Roy Chand, Chancellor****, for his tireless commitment to fostering excellence in teaching and research at Jain (Deemed-to-be-University). I am also profoundly grateful to the honorable* ***Vice Chancellor, Dr. Raj Singh, and Dr. Dinesh Nilkant, Pro Vice Chancellor****, for their unwavering support. Furthermore, I would like to express my sincere thanks to* ***Dr. Jitendra Kumar Mishra, Registrar****, whose guidance has imparted invaluable qualities and skills that will serve us well in our future endeavors.*

▪

I extend my sincere gratitude to ***Dr. Hariprasad S A, Director*** *of the Faculty of Engineering & Technology,* ***and Dr. Geetha G, Director*** *of the School of Computer Science & Engineering within the Faculty of Engineering & Technology, for their constant encouragement and expert advice. Additionally, I would like to express my appreciation to* ***Dr. Krishnan Batri, Deputy Director (Course and Delivery), and Dr. V. Vivek, Deputy Director (Students & Industry Relations),*** *for their invaluable contributions and support throughout this project.*

It is a matter of immense pleasure to express my sincere thanks to ***Dr. Aditya Pai , Program Head Artificial Intelligence and Data Engineering, Computer Science and Engineering****, School of Computer Science & Engineering Faculty of Engineering & Technology for providing right academic guidance that made my task possible.*

I would like to thank our guide ***Mr. Akash Das AVP and Project Manager of Futurense Technologies****, for sparing his valuable time to extend help in every step of my work, which paved the way for smooth progress and fruitful culmination of the project.*

I would like to thank our Project Coordinator ***Mr. Arnab Roy****, and all the staff members of Futurense Technologies for their support.*

I am also grateful to my family and friends who provided me with every requirement throughout the course.

I would like to thank one and all who directly or indirectly helped me in completing the work successfully.

Signature of Student(s)

ABSTRACT

The 120 years of Olympic data from 1896 to 2016 are thoroughly analysed in this paper, with particular attention paid to medal distributions, athlete demographics, historical trends, and their effects on the dynamics of Olympic performance. We investigate patterns including the number of participants over time, the split between male and female athletes, and the gender-specific distribution of weight and height among athletes using a variety of analytical tools. We also look at the relationship between physical characteristics such as weight and height, the age distribution of players in various sports, and patterns of engagement among older age groups.

In addition, our research explores patterns of medal distribution among nations, highlighting the best-performing countries and long-term trends in medal acquisition. We look into the popularity of particular sports, patterns in sports engagement bygender, and distinctions between the Olympics in the summer and winter. We also look at the social and economic aspects of Olympic performance, such as how hosting an Olympics affects a nation's chances of winning medals.

Our research illuminates the history of the Olympic Games and offers athletes, coaches, decision-makers, and scholars insightful information. We emphasise how crucial it is to take into account a number of variables when analysing Olympic competition and performance dynamics, such as medal distributions, historical patterns, and athlete demographics. In order to deepen our understanding of sports dynamics and support evidence-based decision-making in the context of Olympic competition, we conclude by outlining potential directions for future research, such as the investigation of sophisticated predictive modelling approaches and the incorporation of socioeconomic variables into Olympic analysis.

TABLE OF CONTENTS

Chapter 1

1. INTRODUCTION

1.1 Background & Motivation

1.2 Objective

1.3 Delimitation of research

1.4 Benefits of research

Chapter 2

2. LITERATURE SURVEY

2.1 Literature Review

2.2 Inferences Drawn from Literature Review

Chapter 3

3. PROBLEM FORMULATION AND PROPOSED WORK

3.1 Introduction

3.2 Problem Statement

3.3 System Architecture /Model

3.4 Proposed Algorithms

3.5 Proposed Work

4. IMPLEMENTATION

4.1 Software Implementation

4.2 INSIGHTS

Chapter 5

5. RESULTS AND DISCUSSION

CONCLUSIONS AND FUTURE SCOPE

REFERENCES (IEEE FORMAT

APPENDICES

APPENDIX – I

APPENDIX – II

CHAPTER 1

1.INTRODUCTION

1.1 Background & Motivation

The Olympics: Where Glory Meets Data Analysis

"The Olympics" bring together athletes from all over the world, showcasing peak human performance and fostering international unity. The Olympic Games being a pinnacle of international athletic competition, have enthralled spectators for more than a century. Beyond the exhilaration of winning and the sorrow of losing, the Olympics provide an abundance of information detailing the development of physical ability, trends in participation, and the shifting landscape of international sports.

The motivation behind this endeavour stems from unveiling the stories within the data. Beyond the cheers of the crowd and the thrill of victory lies a hidden world: A universe of data waiting to be explored. Today, data analysis plays a crucial role in the Olympics, transforming our understanding of the Games and shaping their future and the goal of this project is to reveal the narratives that are concealed inside these numbers, spanning over 120 years of Olympic history.

This project's importance rests in unlocking hidden stories within the Games. We can analyze participation trends, identify training strategies leading to success, and predict future sports popularity. Imagine uncovering trends in athlete demographics, identifying training methods that lead to success, or even predicting future sports stars. Data analysis empowers athletes, coaches, and organizations to optimize performance, allocate resources, and understand the link between a nation's development and its Olympic achievements. By weaving these insights into the narrative, we gain a richer appreciation for the Games and the ongoing pursuit of excellence on the world stage.

However, there is more to this final project report than merely findings to provide. Its goal is to shed light on the underlying patterns that influence the Olympics in order to provide a more thorough comprehension of this international event.

1.2 Objectives

- Understanding Athlete Demographics and Participation:
    - Analyze the number of athletes participating in the Olympics across different editions.
    - Explore participation trends based on sex (e.g., has female participation increased over time?).
    - Investigate age distribution of athletes across different sports or Olympic Games.
    - Analyze the relationship between athlete age, height, and weight across sports or medal winners.

- Examining Medal Distribution:
    - Identify the total number of medals awarded across all Olympic Games.
    - Analyze the distribution of medals (Gold, Silver, Bronze) across different editions or sports.
    - Investigate which countries (NOC) have won the most medals overall or in specific sports.
    - Understand if there are any correlations between a country's participation numbers and its medal count.

- Identifying Trends in Sports:
    - Determine the total number of sports represented in the Olympics data.
    - Analyze the popularity of different sports based on athlete participation.
    - Explore trends in participation for specific sports across different Olympic Games.
    - Identify emerging or declining sports based on participation data.

- Comparing Summer vs Winter Olympics:
    - Analyze participation and medal distribution differences between Summer and Winter Olympic Games.
    - Investigate the most popular sports in each category (Summer vs Winter).
    - Compare age and physical attribute (height, weight) distributions between Summer and Winter athletes.

- Social and Economic Factors:
    - Identifying if a being a host country influences the winning rate.
    - Female athlete participation rates across different sports and Olympic editions.
    - Examining the potential link between gender equality and female participation across countries.
    - Analyzing how a country's dominant sports or successful athletes contribute to its national identity considering factors like Historical Success and Cultural Significance of sports.

- Predictive Analysis:
    - To identify emerging sports and participation trends, allowing for strategic planning and investment.
    - To build models to predict medal winners in future Olympics or to predict the sport played by an Athlete from their physical attributes.

1.3 Delimitation of Research

- Scope Restriction: The analysis is constrained to predefined objectives in accordance to the dataset which may potentially overlook valuable insights that can be gained from exploring additional datasets.

- Limitation of Data: This project will only focus on data that is publicly available related to Olympics history limiting the analysis to publicly available datasets related to Olympic history. This restriction may limit access to certain datasets curated by private organizations or Olympic committees.

- Limited attribute coverage: While available attributes provide valuable insights, they do not encompass all factors influencing athlete performance or Olympic outcomes. Factors such as training regimes or injury history, which certainly impact on athlete performance are not captured in the dataset which hinders our analysis as these factors may provide additional insights.

- Limitations of Statistical Analysis: The smaller number of Olympic years relative to the total timeframe limits the statistical power of the analysis, potentially reducing the reliability and generalizability of findings related to Olympic participation, performance trends and medal tally.

- Interruptions in Data Continuity: The occurrence of World War I and World War II resulted in interruptions to the Olympic Games, impacting the continuity and completeness of the dataset. These significant historical events led to the cancellation of several editions of the Olympics, thereby limiting the availability of data for certain time periods.

- Temporal Discontinuity: The challenges faced while assessing continuous trends or changes in Olympic related data over time due to intermittent occurrence of Olympic Games every 2-4 years.

- Medal Analysis Limitations: The lack of information regarding the events' level of competition or the criteria used to award medals limits the depth of analysis available on medal distribution. This information is essential for a comprehensive analysis of medal distribution patterns and factors influencing medal success.

- Sports Data limited to Historical Context: The analysis for trends in sports is constrained as the historical data on changes in sports rules, equipment, or training methodologies. This limits the analysis of long-term trends in sports participation and performance.

- Demographic Oversimplification: While analyzing participation by sex and age provides valuable insights, it may oversimplify the diversity of athlete demographics by overlooking factors such as ethnicity, nationality and socioeconomic background which can significantly impact participation trends.

- Limited Evolutionary Insights: The delayed inception of the Winter Olympics limits the ability to conduct comprehensive analyses between Summer and Winter Olympics, potentially omit significant differences and similarities related to sports dynamics.

1.4 Benefits of Research

- Historical Insights: Examining and interpreting data from the past 120 Olympics offers important insights into how the Olympic Games have changed historically, including patterns in medal distribution, athlete participation, and sports dynamics across time.

- Data-driven Decision Making: By leveraging Python programming and data analysis techniques, researchers can uncover patterns, correlations, and trends within the Olympics dataset, enabling informed decision-making for athletes, coaches, sports organizations, and policymakers.

- Performance Optimization: The project can identify factors that contribute to success in Olympic competitions by thoroughly analyzing athlete performance metrics like age, height, weight, and participation in events. This information will inform strategies for athlete preparation, training, and performance optimization.

- Educational Resource: By offering approachable insights and techniques for additional investigation and learning, the project acts as an educational resource for researchers, students, and enthusiasts interested in Olympic history, sports analytics, and data science.

- Public Engagement: The project piques the interest and involvement of the general public, the media, and sports fans. It also initiates conversations about the Olympic legacy, individual athlete accomplishments, and the wider cultural, health, and well-being effects of sports.

The project strikes a balance between its delimitations and benefits by transparently communicating constraints such as data availability and demographic oversimplification, while implementing mitigation strategies to enhance reliability. By embracing continuous improvement, the project remains adaptable and ensures its impact and relevance in the field of sports analytics and Olympic studies.

Chapter 2

2. Review of Literature

2.1 Literature Review

The Olympic Games, spanning over 120 years of history, stand as a monumental testament to human athleticism, competition, and unity. Scholars and researchers have delved into the vast repository of Olympic data, seeking to uncover patterns, trends, and insights that illuminate the evolution of this global sporting spectacle. The following literature review provides an overview of studies that have explored the attributes encompassed within the dataset of Olympic athletes, ranging from demographic information to performance metrics.

- Yamunathangam, Kirthicka, and Shahanas Parveen (May 2019): Yamunathangam, Kirthicka, and Shahanas Parveen's study represents a pioneering effort in the realm of Olympic data analysis. Their research focuses on leveraging exploratory data analysis techniques to delve into the intricacies of athlete performance across various Olympic Games. By examining attributes such as age, height, weight, and team affiliations, the researchers construct a comprehensive picture of the demographic and competitive landscape of the Olympics. Through data visualization techniques, they provide users with intuitive insights into the historical trends and patterns that have shaped the Games over the decades. However, their analysis falls short in forecasting the likelihood of victory or predicting winning medal constellations, indicating avenues for further research in predictive modeling.

- Dominik Schreger, Wunderlich, Limas, and Sacha Schmidt (December 2020): Schreger, Wunderlich, Limas, and Schmidt's study shifts the focus towards predictive modeling and forecasting within the realm of Olympic data analysis. Employing the Random Forest Algorithm, the researchers aim to predict the medal counts of countries in future Olympic Games. Central to their analysis are attributes such as age, height, weight, and past performance data, which serve as inputs for training the predictive model. By assessing success drivers and benchmarking team performance, the study offers valuable insights into the factors influencing Olympic success. However, the absence of a novel method to address missing data highlights a potential area for methodological refinement in future research endeavors.

- Rahul Pradhan (January 2021): Pradhan's research endeavors to unravel the historical tapestry of the Olympic Games through a nuanced examination of athlete attributes and their relationships. By employing exploratory data analysis techniques, Pradhan seeks to visualize the intricate web of factors shaping Olympic success, ranging from demographic variables to sporting disciplines. Attributes such as age, height, weight, and participation in specific events are scrutinized to discern patterns and trends across different years, seasons, and host cities. While Pradhan's study offers valuable insights into the dynamics of Olympic competition, the exploration of alternative machine learning techniques for data representation presents a promising avenue for future research exploration.

- In summary, the literature review underscores the multifaceted nature of Olympic data analysis, encompassing demographic attributes, performance metrics, and predictive modeling techniques. As scholars continue to unravel the complexities of this rich dataset, opportunities abound for further exploration, innovation, and discovery within the realm of Olympic research.

- Ashay Maheshwari(2018): Some interesting insights on the data we have available, like say person who won most number of golds in olympic history, number of countries participated each year and what not.
- Learning purpose - pandas, matplotlib and seaborn libraries were used to analyse the data and provide us an interesting use case to apply these skills

- GABRIEL PREDA (2021): Plotly tutorial - 120 years of Olympic games.
  Here "GABRIEL PREDA" has explained about the analysis and plot chart types and functions. Two types of Content lists are used: for the analysis of the dataset, we use the Analysis content list. For the Plotly features, he used the Plotly chart types and functions content list.Analysis by Participationand etc.

2.2 Inferences Drawn from Literature Review

The Value of Data Visualisation Techniques in Interpreting Past Olympic Performance Trends:
Because the amount and complexity of the data involved in an Olympic performance are so great, data visualisation tools are essential to understanding past performance trends. Through the use of graphical representations like heatmaps, graphs, and charts, researchers are able to condense large datasets into easily understood visual stories.

The Potential for Predictive Modelling in Medal Outcome Forecasting Using Machine Learning Algorithms, Like Random Forest:
The use of machine learning algorithms, especially ensemble techniques such as Random Forest, has enormous promise for predictive modelling in Olympic medal projections. These algorithms are excellent at finding intricate links and patterns in data, which makes them a good choice for forecasting how athletes and nations will fare in the Games.

The necessity of fixing missing data and enhancing performance models on a constant basis in Olympic analysis
Ensuring the quality and dependability of performance models in Olympic analysis requires addressing missing data. Absence of data can induce biases, distort outcomes, and call into question the reliability of inferences made using analytical models. In order to evaluate the effect of missingness on model outcomes, researchers must use reliable methodologies for addressing missing data, such as imputation techniques, data augmentation, or sensitivity studies.

The possibility of integrating various analytical methods, such as exploratory data analysis and machine learning, for comprehensive insights into Olympic performance dynamics.

These inferences suggest a multidimensional approach to analyzing Olympic Games data, combining the strengths of different methodologies to achieve more robust and insightful conclusions.

The study showcases the significance of exploratory data analysis techniques in uncovering historical trends and patterns in Olympic athlete performance. However, it highlights the need for further research in predictive modeling to forecast victory likelihood and medal outcomes accurately.

This research underscores the importance of predictive modeling and forecasting in understanding the factors influencing Olympic success. While the study provides valuable insights into success drivers using the Random Forest Algorithm, it indicates the necessity for addressing missing data to enhance methodological robustness.

Pradhan's study emphasizes the nuanced examination of athlete attributes and their relationships to unravel the dynamics of Olympic competition. It suggests the exploration of alternative machine learning techniques for data representation to further enrich insights into Olympic performance.

This resource provides a practical demonstration of data analysis using libraries like pandas, numpy, matplotlib, and seaborn. It offers insights into various aspects of Olympic data, such as the most successful athletes and the participation of countries over the years, serving as a valuable learning resource for data analysis enthusiasts.

The tutorial by GABRIEL PREDA delves into the utilization of Plotly for analyzing 120 years of Olympic Games data. It provides insights into different types of analysis and visualization techniques using Plotly, thereby enhancing understanding and application of advanced charting functionalities.

In summary, the literature review and additional resources collectively highlight the diverse methodologies and tools employed in analyzing Olympic data, ranging from exploratory data analysis to predictive modeling and advanced visualization techniques. These studies offer valuable insights into the evolving landscape of Olympic research and underscore the importance of interdisciplinary approaches in unraveling the complexities of this rich dataset.

CHAPTER 3

3. PROBLEM FORMULATION AND PROPOSED WORK

3.1 Introduction

In this chapter, we delineate the problem formulation and outline the proposed work for our research project in Olympics analytics. Building upon the literature review conducted in Chapter 2, we aim to address key research questions and objectives pertaining to the analysis of the Olympics Dataset. By formulating clear research objectives and delineating the proposed methodology, this chapter sets the foundation for the subsequent chapters of our research.

Specifically, this chapter begins with an introduction to the overarching research problem and provides context for the proposed work in Olympics analytics. We outline the objectives of our research project and discuss the significance of addressing these objectives in the context of Olympics analytics. Furthermore, we provide an overview of the methodology employed in our research, including data collection, analysis techniques, and tools utilized for data processing and visualization.

Through this chapter, we aim to establish a clear framework for our research project, elucidating the scope, objectives, and methodology employed. By defining the problem space and outlining our proposed approach, we set the stage for the subsequent chapters, where we delve into the detailed analysis of Olympics data and derive actionable insights on the Olympics dataset.

3.2 Problem Statement

The Olympic Games, a symbol of unity and athleticism, provide a unique lens through which we can examine global sports participation, cultural dynamics, and athletic achievements. With access to a comprehensive dataset spanning various decades, encompassing diverse sports, and featuring athletes from around the world, our objective is to delve into the intricacies of Olympic history.
Our study aims to address the following key questions:

Historical Evolution: How has the landscape of the Olympic Games evolved over time in terms of participation, representation, and diversity? By analysing trends in athlete demographics (age, sex, nationality) across different epochs, we aim to elucidate the changing dynamics of global sports engagement.

Sporting Trends: What are the dominant sports and events within the Olympic movement, and how have they evolved over the years? By examining patterns in sports popularity, emergence, and decline, we seek to uncover underlying factors driving shifts in athletic preferences and trends.

National Performance: How do different nations fare in terms of Olympic success, and what factors contribute to their performance? Through a comparative analysis of medal

tallies, athlete demographics, and investment in sports infrastructure, we aim to discern patterns of sporting excellence and identify potential determinants of national success.

Gender Equality: To what extent has gender representation and equity been achieved within the Olympic Games? By scrutinizing participation rates, medal distributions, and policy initiatives aimed at promoting gender equality, we aim to evaluate progress towards fostering inclusivity and diversity within the Olympic movement.

Sociocultural Impact: How do the Olympics reflect and influence broader societal trends, including geopolitics, cultural exchange, and globalization? Through qualitative analysis of historical events, controversies, and symbolic moments, we aim to explore the multifaceted role of the Olympics as a global platform for diplomacy, cultural expression, and social change.

By addressing these questions, our study seeks to provide a comprehensive understanding of the historical trajectory, societal impact, and cultural significance of the Olympic Games. Through data-driven analysis and contextual interpretation, we aim to contribute valuable insights into the evolving nature of global sports culture and the enduring legacy of the Olympic movement.

3.3 System Architecture /Model

The proposed system architecture/model for our Olympics analytics research project comprises several interconnected components designed to facilitate data collection, processing, analysis, and visualization. The architecture/model is structured to enable comprehensive analysis of Olympics dataset. The key components of the system architecture/model are as follows-

- Data Collection: Data collection involves gathering Olympics data from reliable sources such as the official Olympics website, sports statistics websites, and data Automated data scraping tools or APIs may be utilized to retrieve structured data from online sources.
- Data Pre-processing: The collected data undergoes pre-processing to clean, transform, and standardize the dataset for analysis. This includes handling missing values, data normalization, and feature engineering to extract relevant variables for analysis Data

pre-processing techniques ensure the integrity and quality of the dataset before further analysis.

- Data Storage: Processed data is stored in a centralized database or data warehouse for efficient storage and retrieval. The database may be organized into structured tables or files, enabling fast access to specific datasets and facilitating data manipulation for analysis purposes. Cloud-based storage solutions or relational databases may be utilized for scalable and reliable data storage.
- Data Analysis. Data analysis involves applying statistical techniques, machine learning algorithms, and data visualization methods to extract insights from the Olympics dataset. Statistical analysis may include calculating averages, medal distributions player performances and match outcomes. Machine learning models may be employed for predictive modelling, clustering, or classification tasks to identify patterns and trends in the data.
- Visualization and Reporting Visualizations such as charts, graphs, and dashboards are generated to present the insights derived from the analysis in a clear and intuitive manner. Visualization tools such as matplotlib, seaborn, or Tableau may be utilized to create interactive visualizations that facilitate exploration and interpretation of the data.

The proposed system architecture/model provides a structured framework for conducting comprehensive analysis of Olympics dataset. By leveraging advanced data analytics techniques and visualization tools, the architecture/model enables one to gain deeper insights into player performances, team strategies, match dynamics, and external factors influencing match outcomes such as GDP, social factors etc.

3.4 Proposed Algorithms

- Descriptive Statistics: Utilize descriptive statistics such as averages, totals, medal distributions, and correlations to summarize and interpret match data of different sports. This can involve calculating various summary statistics depending upon the sport to gain insights into player and team performances.
- Trend Analysis: Conduct trend analysis to identify patterns and trends in various sports' match data over time. This may involve analysing season- wise

performance trends, venue-wise performance variations, or the impact of external factors such as weather conditions on match outcomes.

- Comparative Analysis: Compare performance metrics across different teams, players, seasons, and match conditions to identify relative strengths and weaknesses. This can help in benchmarking performances and assessing the effectiveness of different strategies employed by teams.
- Match Outcome Analysis: Analyse factors contributing to match outcomes, such as toss decisions, squad lineups, playing strategies, teammate synergy, individual performances, etc. By examining the relationship between these factors and match results, you can gain insights into the determinants of success across various sports.
- Visualization Techniques: Use data visualization techniques such as charts, graphs, and heatmaps to visualize match data depending upon the sport and identify patterns visually. This can aid in communicating insights effectively understanding performance and results more effectively.
- Exploratory Data Analysis (EDA): Conduct exploratory data analysis to explore and understand the underlying structure of match data. This may involve data cleaning, outlier detection, and data transformation to prepare the dataset for analysis.

- Qualitative Analysis: Supplement quantitative analysis with qualitative insights obtained from match commentaries, expert opinions, and post- match analyses. Qualitative analysis can provide context and nuance to quantitative findings, enhancing the overall understanding of cricket match dynamics.

By employing these approaches, you can conduct meaningful analysis of any match data irrespective of the sport without relying heavily on specific algorithms. This allows for a more flexible and exploratory approach to sports analytics, focusing on extracting insights from data using a variety of statistical and analytical techniques.

3.5 Proposed Work

The Olympic Games stand as a testament to human athleticism, international camaraderie, and cultural exchange. With a rich history spanning over a century, the

Olympics offer a treasure trove of data encompassing athlete demographics, sporting events, national performances, and societal impact. In this proposed work, we aim to harness advanced analytical techniques to unlock hidden insights from the vast reservoir of Olympic data, shedding light on historical trends, athlete performance dynamics, and the broader societal implications of the Olympic movement.

- Data Collection and Preprocessing:
  The first phase of our proposed work involves the comprehensive collection and preprocessing of Olympic data. Leveraging publicly available datasets and archival sources, we will compile a structured repository containing information on athletes, events, participating nations, and historical context. Data preprocessing tasks will include cleaning, standardization, and integration to ensure the consistency and integrity of the dataset across various dimensions such as time periods, sports disciplines, and event categories.
- Exploratory Data Analysis (EDA):
  Once the dataset is curated and standardized, we will conduct exploratory data analysis to gain initial insights into the underlying patterns and trends within the Olympic data. Through descriptive statistics, data visualization techniques, and interactive dashboards, we will examine key metrics such as athlete demographics, medal distributions, sports popularity, and temporal variations in Olympic participation. EDA will serve as a foundation for hypothesis generation and guide subsequent stages of our analytical framework.
- Clustering Analysis:
  Building upon the insights garnered from EDA, we will employ clustering analysis techniques to uncover latent patterns and groupings within the Olympic dataset. Using algorithms such as K-Means and Hierarchical Clustering, we will segment athletes, events, and nations based on shared characteristics such as demographic profiles, performance metrics, and historical contexts. Clustering analysis will facilitate the identification of distinct clusters within the data, enabling deeper exploration into the factors driving diversity and differentiation within the Olympic ecosystem.
- Predictive Modelling:
  In parallel with clustering analysis, we will develop predictive models to forecast future trends and outcomes within the Olympic domain. Regression analysis will be employed to model the relationship between athlete attributes (e.g., age, height,

weight) and performance metrics (e.g., medal count, event participation), allowing us to quantify the impact of demographic factors on athletic success. Additionally, classification algorithms such as Decision Trees and Random Forests will enable us to classify athletes into performance categories (e.g., medallists vs. non-medallists) and identify key predictors of Olympic success.

- Network Analysis:
  Another key aspect of our proposed work involves the application of network analysis techniques to examine the interconnectedness and dynamics within the Olympic ecosystem. Social Network Analysis (SNA) will be used to analyse the relationships among athletes, teams, and nations based on co-participation in events and shared characteristics. Additionally, event co-occurrence networks will be constructed to visualize the clustering of related sports disciplines and uncover emerging trends within the Olympic program.
- Text Mining and Sentiment Analysis:
  To complement quantitative analyses, we will leverage text mining and sentiment analysis techniques to extract insights from textual data sources related to the Olympics. By analyzing athlete biographies, event descriptions, media coverage, and official reports, we aim to identify prevailing narratives, sentiment trends, and thematic patterns surrounding the Olympic Games. Topic modeling algorithms such as Latent Dirichlet Allocation (LDA) will enable us to uncover latent topics and themes within textual data, providing a qualitative perspective on the societal impact and cultural significance of the Olympics.

Conclusion:

In summary, our proposed work offers a holistic approach to analysing Olympic data, encompassing a diverse array of advanced analytical techniques tailored to the unique challenges and opportunities presented by the Olympic domain. Through data-driven exploration, predictive modelling, network analysis, and text mining, we aim to unravel the intricacies of Olympic history, athlete performance dynamics, and the broader societal implications of the Olympic movement. By shedding light on hidden patterns and trends within the data, our research endeavours to contribute valuable

insights to the fields of sports analytics, cultural studies, and data science, enriching our understanding of one of the world's most celebrated sporting events.

Chapter 4

4.1 Implementation

4.1 Software algorithm

A comprehensive approach utilising a range of tools, technologies, and methodologies is needed to implement software solutions for the analysis of a dataset that spans 120 years of Olympic data from 1896 to 2016. The dataset includes attributes such as Name, Sex, Age, Height, Weight, Team, NOC (National Olympic Committee), Games, Year, Season, City, Sport, Event, and Medal. An overview of the software implementations specifically designed for this dataset is provided below:

Data Preprocessing and Cleaning: To address missing values, inconsistencies, and outliers in the dataset, data processing pipelines are put into place. Before analysing data, it is necessary to ensure its consistency and quality through the development of scripts and queries using programming languages like Python.

Exploratory Data Analysis (EDA): EDA methods are used to examine and show the dataset in order to find trends, patterns, and connections between different variables. Researchers can learn more about medal distributions, athlete demographics, and historical patterns by utilising Python libraries and frameworks like NumPy, Pandas, Matplotlib, Seaborn, and Plotly. These tools make data processing, visualisation, and statistical analysis easier.

Insights Gained

Analysis by Participation; What is the Athlete Participation per Olympic year,What is the proportion of Male vs Female participation & did it increase over time,The Olympic Games were canceled in 1916, 1940, and 1944. Were these cancellations connected to any historical events.

Analysis by Age; What is the age dynamics across Sports,What is the influence of age on medal type.

Analysis by Medal; What is the Medal Distribution by country,What is the distribution of medals by each year and sport,What is the distribution of medals for each Gender.

Analysis by Sport :What are the most popular sports by participation,What is the correlation of height and weight for each sport,What is most and least played sport by women over the years.

Analysis by Season : What is the number of sports played by each edition,Popular sport per edition based on Athlete participation,Which countries participated most in each of the seasons.

Social Factors : What are the countries having the best male female ratio,Does hosting the Olympics improve Performance,What is the gender distribution in top countries.

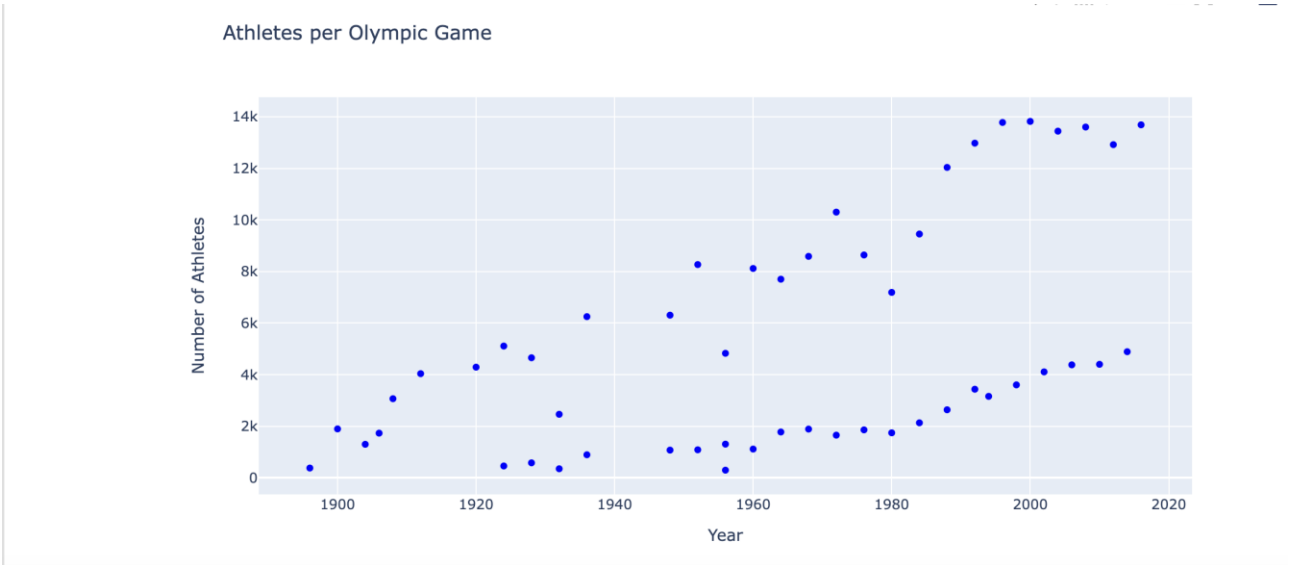
Predictive Analysis :What are the most popular sports by participation,What is the correlation of height and weight for each sport,What is most and least played sport by women over the years.

Reporting and Visualisation:
In order to communicate data-driven insights, interactive dashboards, charts, and reports are created using visualisation tools and libraries. Researchers can create and implement user-friendly interfaces for analysing Olympic data by using frameworks like matplotlib, pyplot, and seaborn. This promotes data-driven decision-making and the sharing of knowledge.
RESULTS AND DISCUSSION .

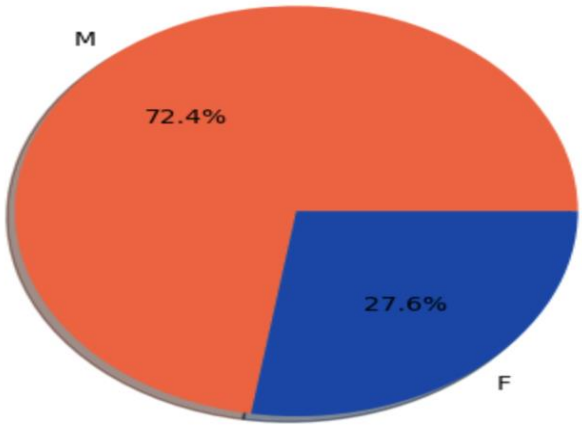
The findings of our examination of the 120 years of Olympic statistics, from 1896 to 2016, are shown in this section. We start out by going over some of the most important conclusions about medal distributions, athlete demographics, and historical trends found in the dataset. We next explore the ramifications of these results, providing insights into the variables affecting Olympic performance dynamics and success in various sports, competitions, and nations. In addition, we talk about the limits of our analysis and possible directions for further study to improve our comprehension of Olympic competitiveness and athlete performance. Here we implement various techniques to identify various trends. Some of them are listed below

How many number of participants participated over the years in every Olympic Game?

The dataset covering the year span from 1896 to 2016 shows a comparatively higher number of participants; nonetheless, the lowest number of participants is 380 (1896) and the largest number is approximately 13.8K.

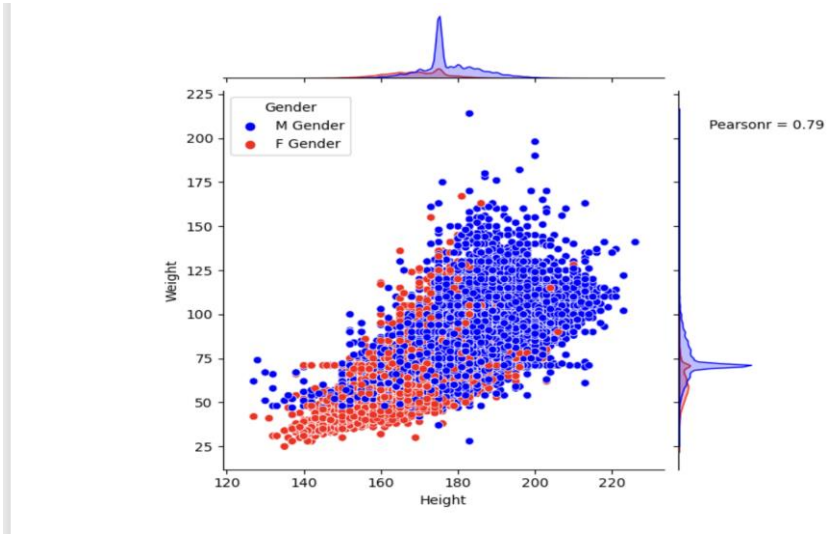


*What is the proportion of male vs female participants in the olympics?
Male participants are more than Female Participants but how do we observe if the number of female participants have increased over the years or not? Let's dive in.

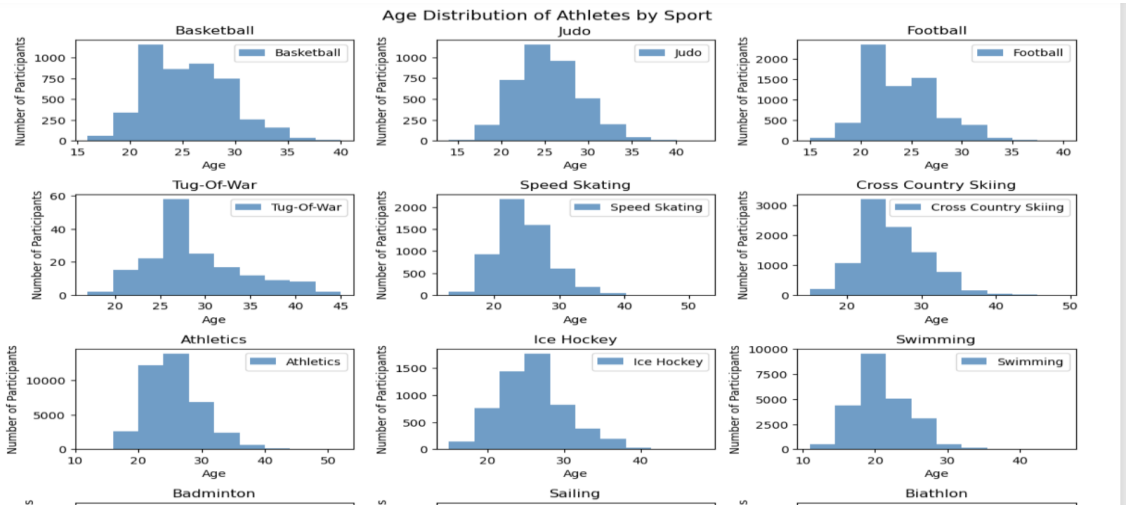


*Understanding the height and weight distribution of Athletes based on gender.
Height Distribution: Women's height increases steadily from 160 to 175 cm, whereas men typically reach a maximum height of 175 cm.
Weight Distribution: Women's weight peaks at 59 kg and 70 kg, whereas men's average weight is 70 kg.

*What is the correlation of attributes like height and weight based on their sex?
The pearson correlation between the height and weight of the athletes is observed to be 0.79 which indicates that it is a strong correlation and hence the weight and height with respect to the sex of atheletes is correlated.The age range of 20 to 30 years old is the common region covered for both genders, with the greatest number of participants from both groups falling within this age range.

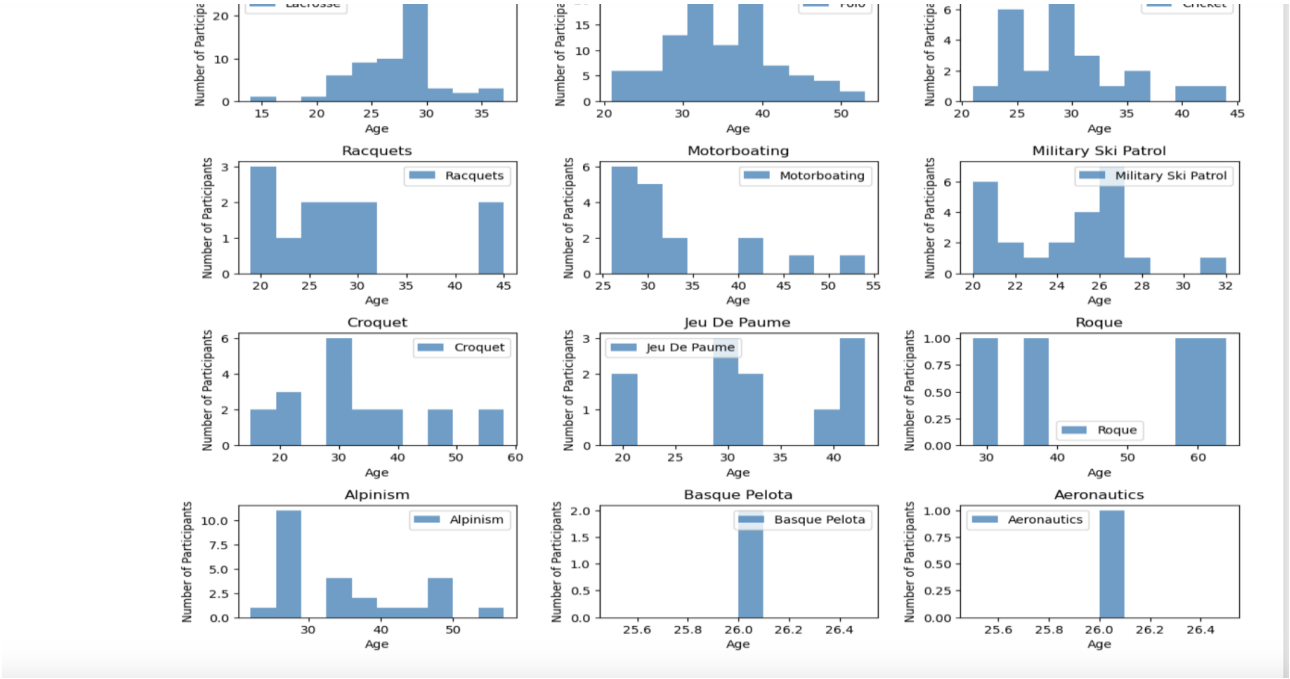


*How is the age distribution among athletes across different Sport?
The age range for the top sports is found to be between 20 and 30 years old, while there are also a significant number of participants in the 40+ age range. The older age groups are observed to be participating in sports that require more mental strength. A deeper comprehension of the age dynamics within each sport category can help sport enthusiasts as well as researchers to identify trends and patterns in participation.Since we know that there is equal number of participation between the younger age groups,

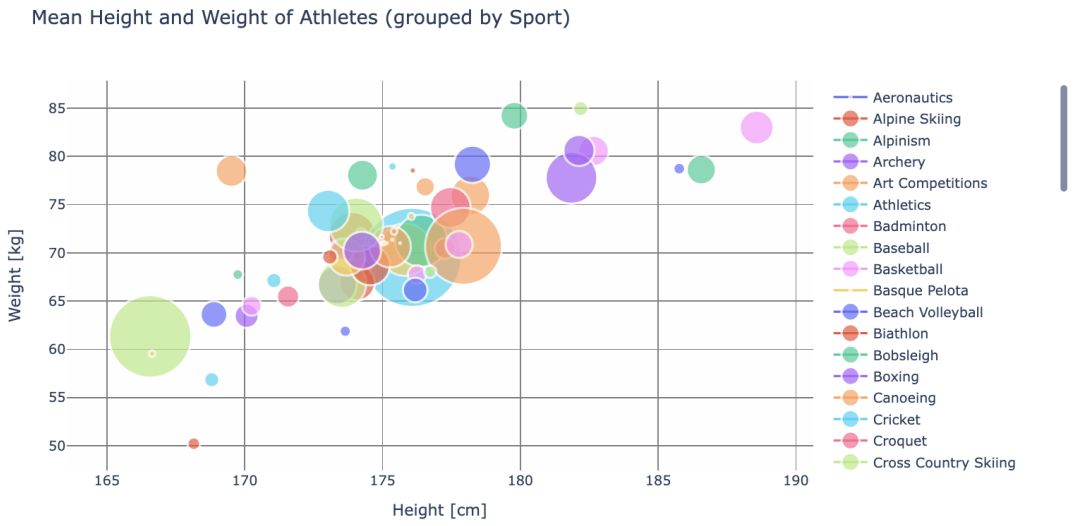


*how do we determine the older age groups participation trends

An interesting observation is that individuals above 40 have been doing in Art competitions,Shooting,Equestrianism, Sailing,and Archery which proves our observation from the previous analysis that they participate in sports that require more mental strength.



*What is the average Height and Weight of Olympiads for each Sport?
Majority of Sports are visibly holding an average weight of around 65-79kgs and average height of around 173-177cms. The plot provides valuable insights into the physical attributes of athletes in various sports which unlocks the athletes participation trends with respect to the physical attributes.

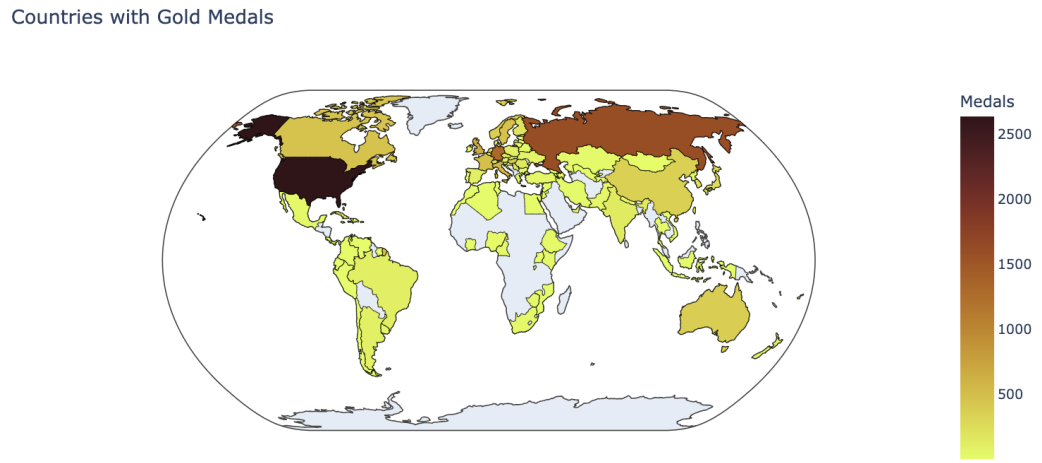


Medal Distribution

Medal Distribution analysis can help provide valuable insights into historical trends, identify areas of dominance, and potentially predict future Olympic performance.

Let's dive into the distribution of medals across countries to understand the top winning countries by each category (Gold,Broze,Silver)

Gold;

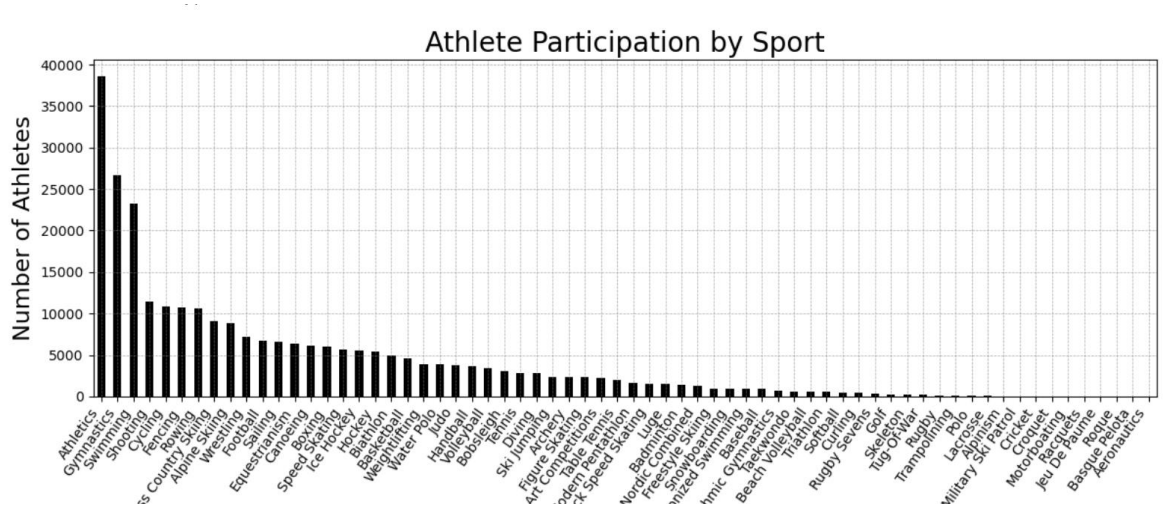


United States, Russia, Germany are the nations with the highest number of gold medals won, while the Northwest region of South America and the African continent have the lowest records
The United States, Russia, and Germany are the nations with the most number of medals in each category, indicating that they have the best chance of winning; in contrast, Southeast Asia, Africa, and the Northwest parts of South America consistently have the fewest medals. If these nations send more participants, it may be possible to lower this in subsequent years and eventually raise the winning rate.

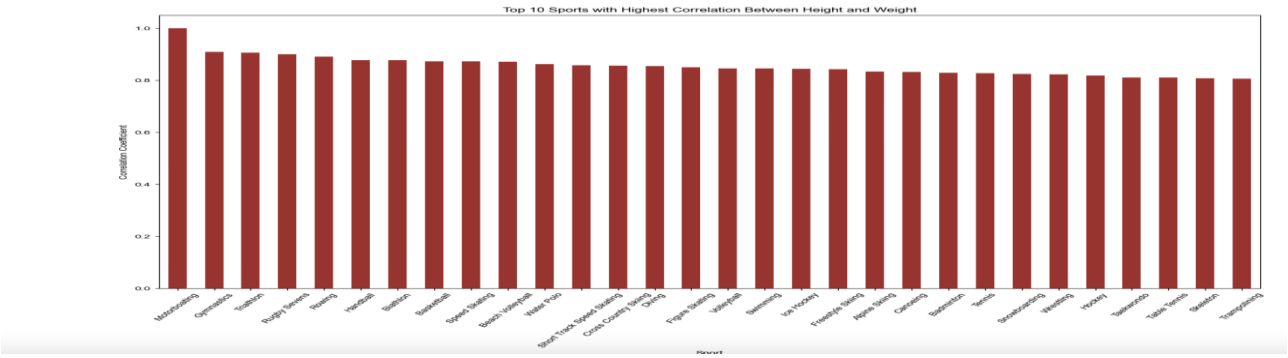
Trends in Sports

By analyzing trends in Sports, one can uncover significant trends in sports played at the Olympic level and gain a deeper understanding of how sports evolve over time.The general

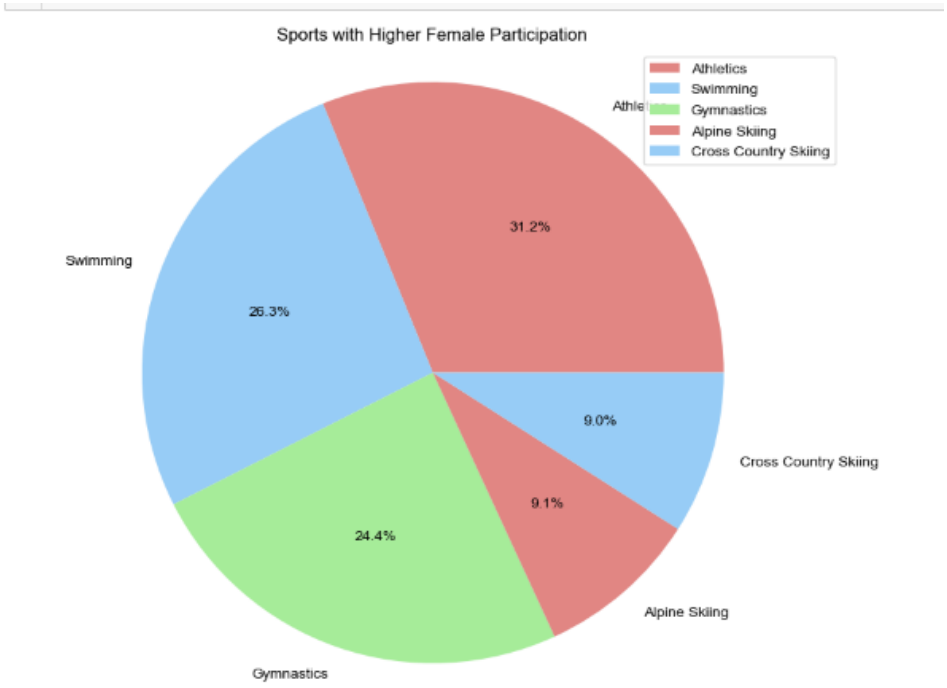
correlation coefficient of height and weight is 0.787 like calculated before. However, this relationship varied significantly among different sports:

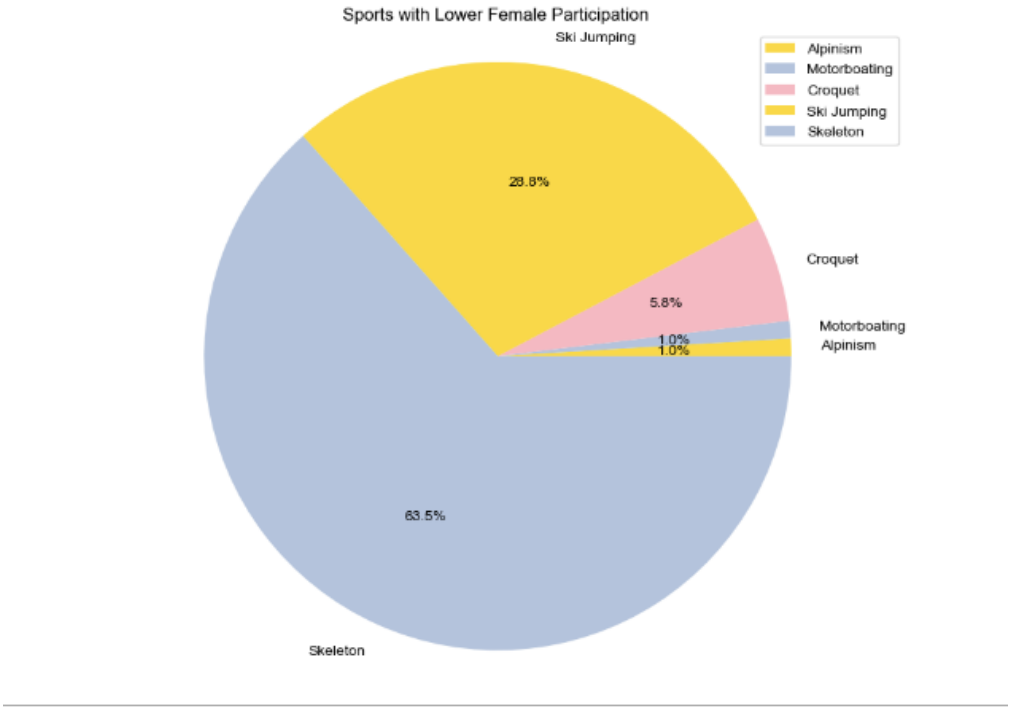


High Correlation: Activities like Rugby Sevens, Gymnastics and Triathlons showed strong correlations, indicating that an athlete's height and weight are closely aligned in these activities.
Low Correlation: Athletes' physical attributes varied more in sports like Tug-Of-War and Ski Jumping, which had lower correlations.
Particular Cases: Because motorboating and aeronautics are two distinct sports due to their specific nature. Hence it is possible that certain correlation patterns are unique.



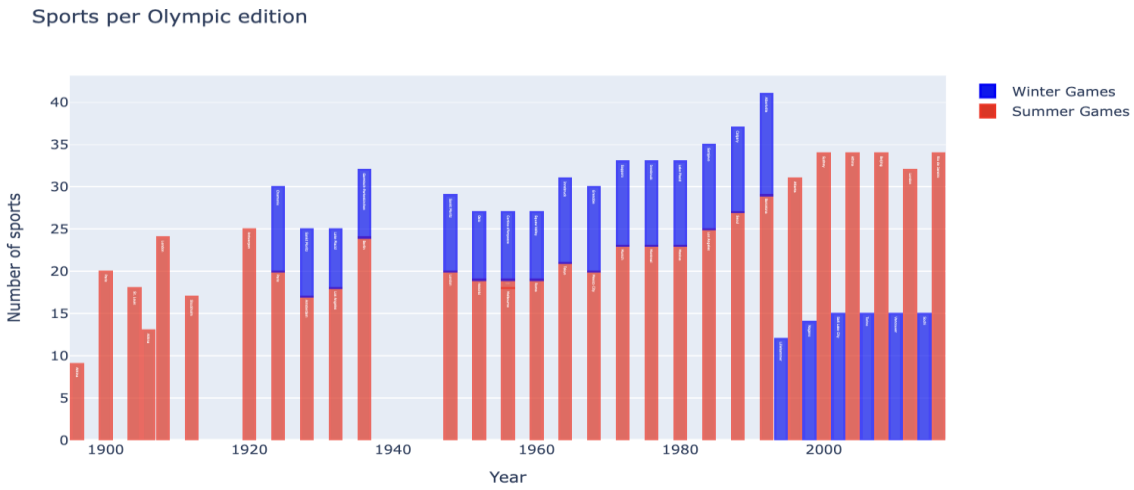
*Women Dominance: What is most and least played sport by women over the years?
Preferred Sports: The sports with the highest female participation rates were gymnastics, swimming, and athletics.
Less Preferred Sports: Women participated at lower rates in motorboating, croquet, and alpinism.



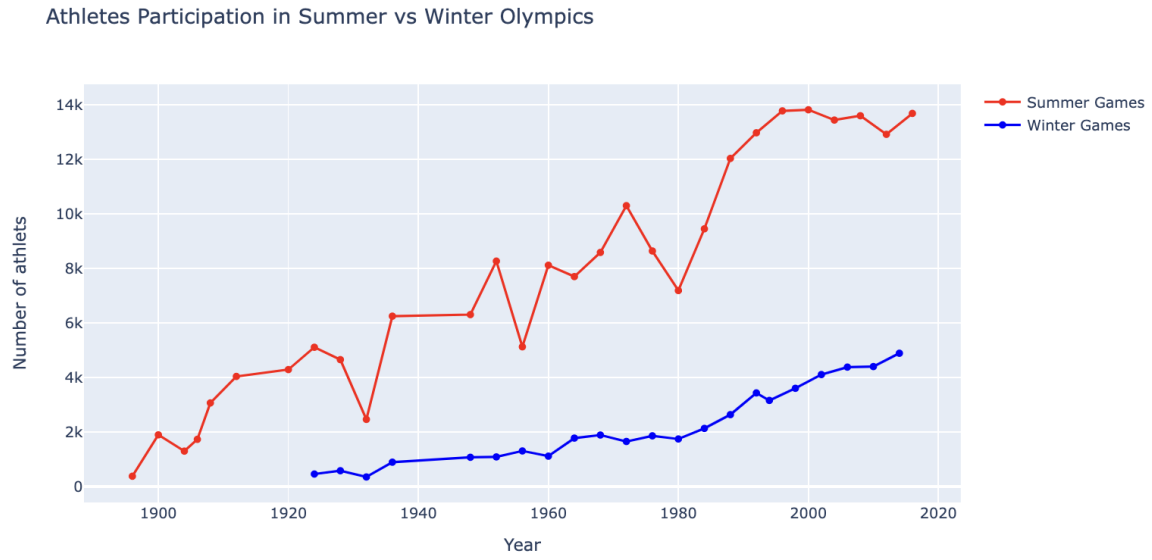


Summer vs Winter Olympics.
*What is the number of sports played per season.

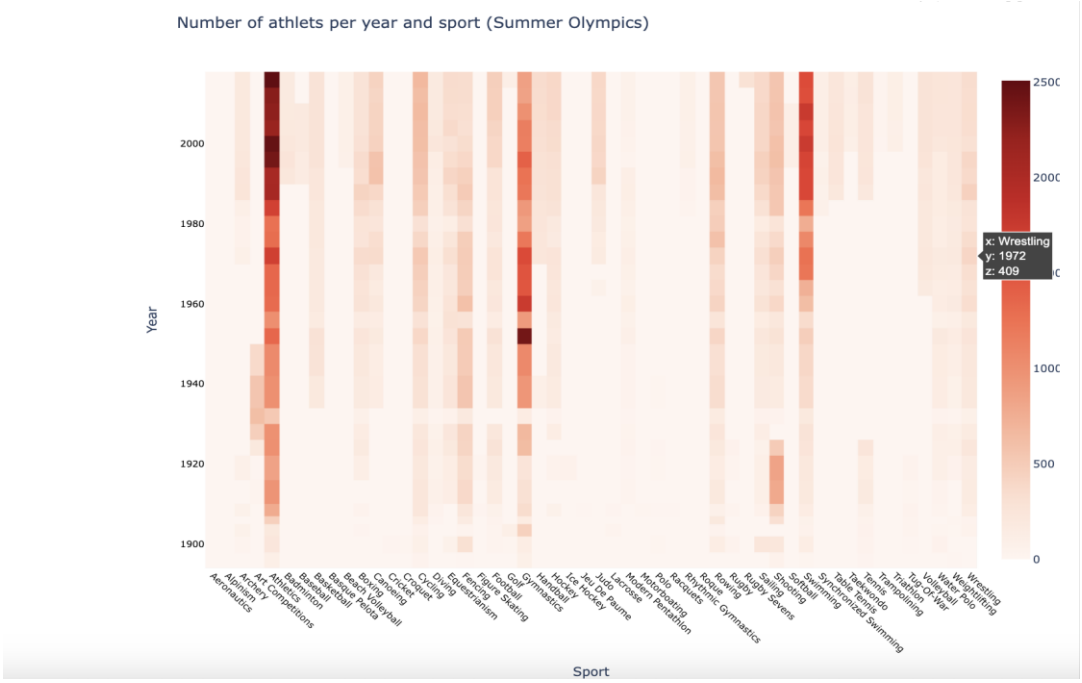
Ten sports were competed at the inaugural Winter Olympics in 1924 in Chamonix. Although the games have since advanced, the number of sports competed in the Winter Olympics is far smaller than in the Summer Olympics

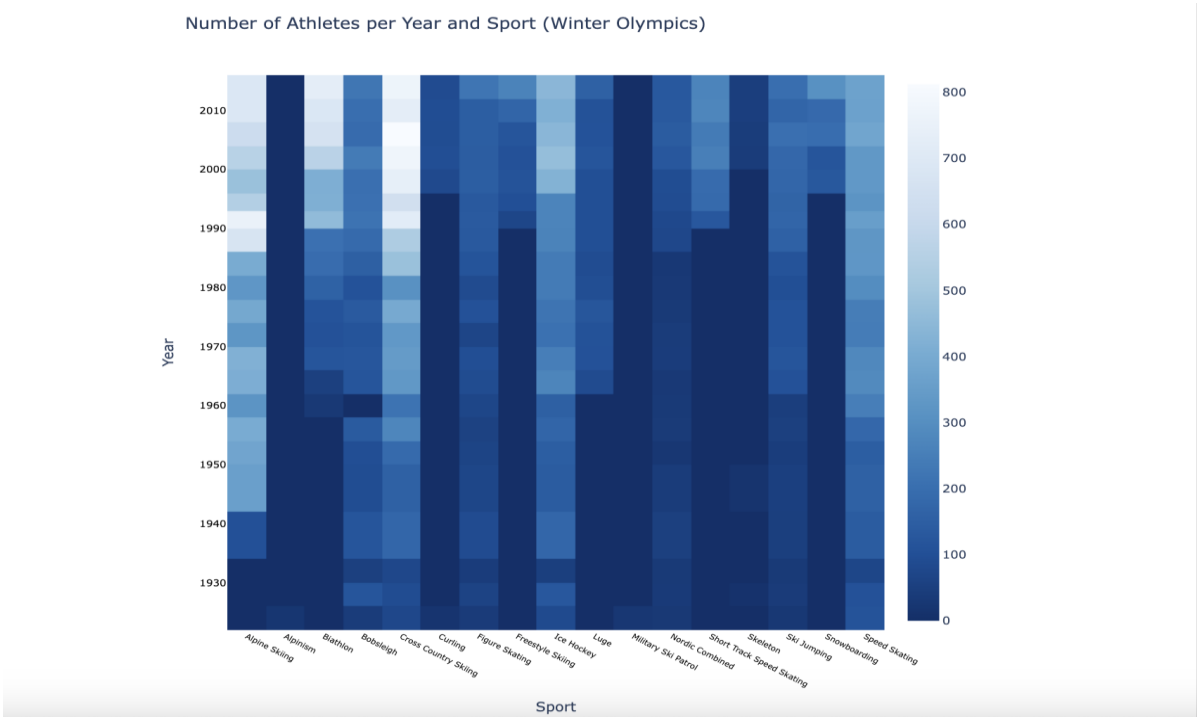


*what is the difference in the participation of athletes in Summer vs Winter Olympics over the years?
By utilizing scatter plots,it provides a clear comparison between the number of athletes involved in the two types of Olympic Games. When comparing the Summer Olympics to the Winter Olympics, we can see that the trend of participation is increasing in Summer Olympics. Additionally, it is noted that while athlete participation in the winter Olympics has been steadily increasing, that of the summer Olympics has also had its significant amount of decreases in participation and hence the graph is not steadily increasing



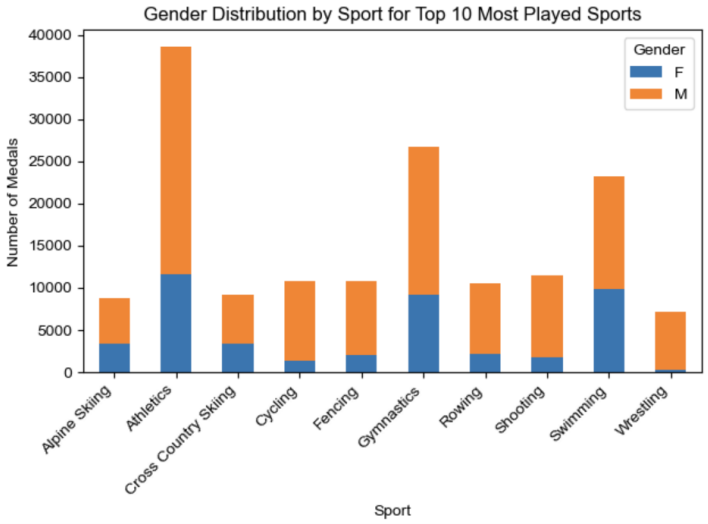
*How many athletes participated in the winter and summer olympics per year Social and Economic Factors.



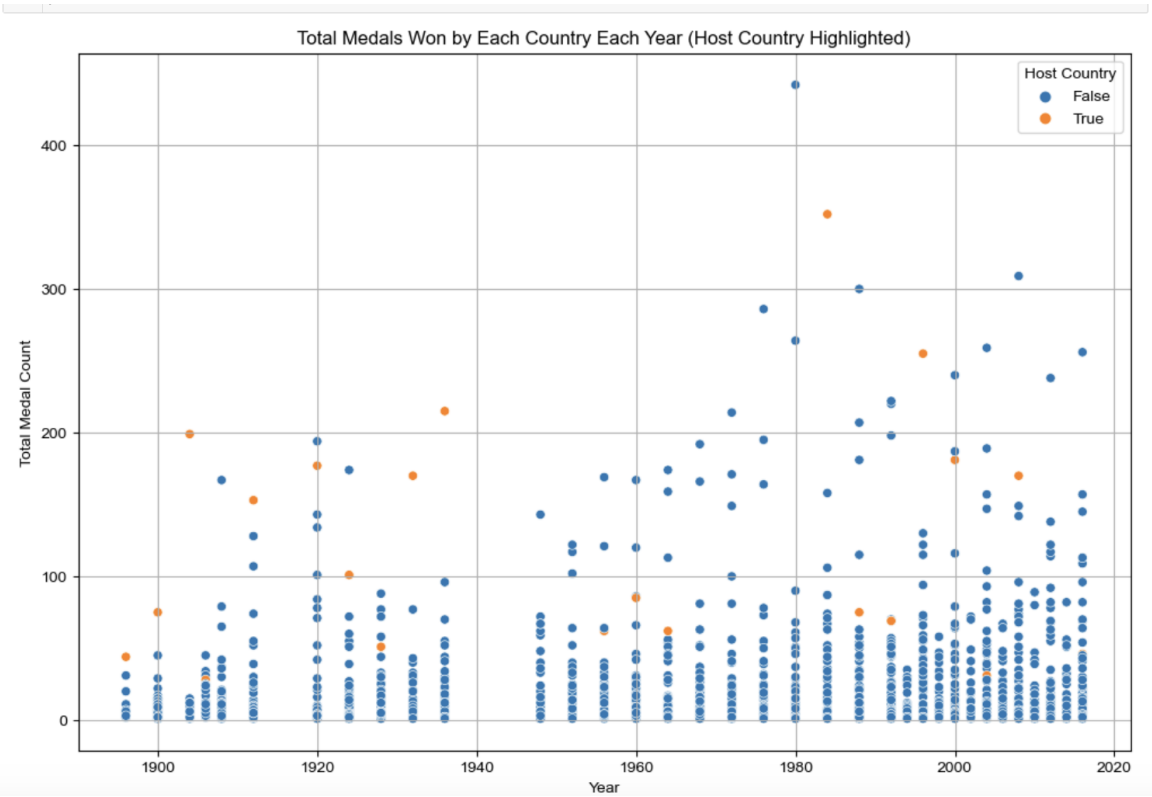


Analyzing social and economic factors can provide valuable insights into a country's Olympic performance. Here's how you can uncover these factors based on our analysis. Countries that have the best male female ratio
and sport.

Countries like China, Guinea-Bissau, Marhsall Islands, North Korea and Palau place the highest in sending equal number of participants with respect to gender.



*Does hosting the olympics improve Performance.
Early Years (Upto 1940): Host nations regularly achieved the top ranks, often taking the top spot.
Mid Years (1945–1980): There was less of a continuous trend, with host nations not always coming in first.
Recent Years (Post 1990): Host countries generally performed well, often ranking in the top 3, with a few exceptions.



These results imply that, although serving as the host nation can boost output, it is not a surefire way to win the most medals. Geopolitical, social, and economic factors,play crucial roles in a nation's Olympic success.

In conclusion, our analysis of 120 years of Olympic data has provided valuable insights into the evolution of the Olympic Games and the factors shaping athletic performance over time. We have identified trends in athlete demographics, medal distributions, and participation patterns across different Olympic Games, seasons, and host cities. However, there remain several avenues for future research, including the exploration of advanced predictive modeling techniques, longitudinal studies of athlete development, and the integration of socio-economic factors into Olympic analysis. By continuing to analyze and interpret Olympic data, researchers can enhance our understanding of sports dynamics, inform evidence-based policy decisions, and inspire future generations of athletes.

REFERENCES

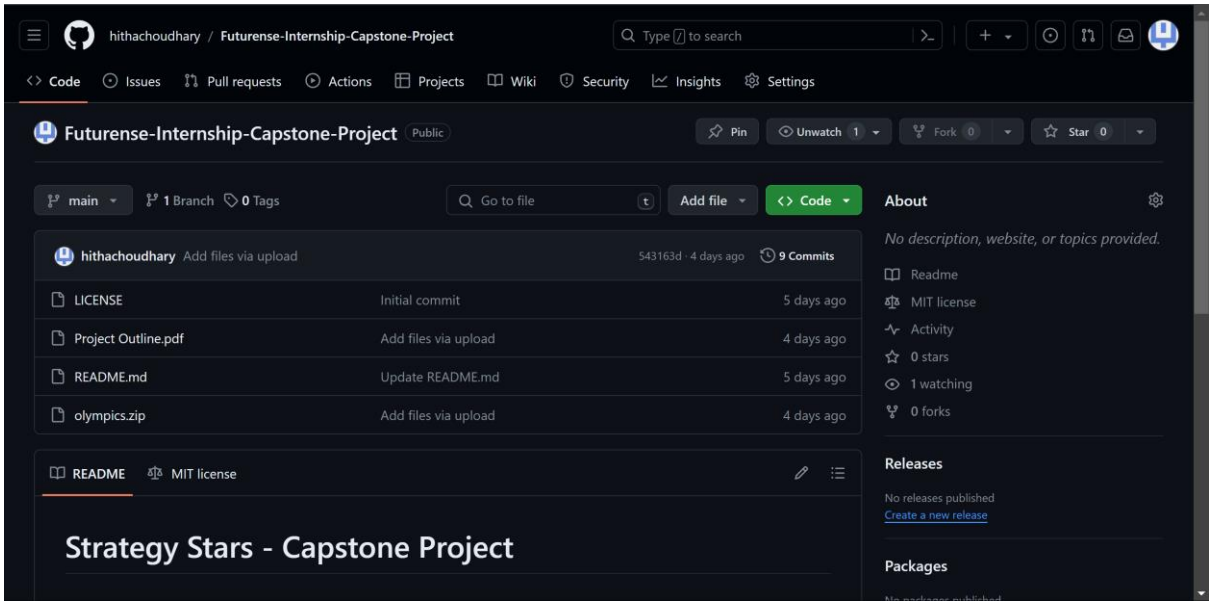
Read:

- https://www.kaggle.com/code/chadalee/olympics-data-cleaning-exploration-prediction/notebook#Can-we-predict-the-medal-tally-of-a-country
- https://www.linkedin.com/pulse/predicting-olympic-medal-winners-holden-price?utm_source=share&utm_medium=member_ios&utm_campaign=share_via
- https://rstudio-pubs-static.s3.amazonaws.com/801501_af7f67d76fe4420baa0e1a36d1432648.html
- https://github.com/hrugved06/Olympics-Medal-Prediction/blob/main/Model/Olympic_Medal_Prediction.ipynb
- https://github.com/alteryx/predict-olympic-medals/tree/master
- https://www.kaggle.com/code/aaronsnowberger/olympics-data-cleaning-exploration-prediction#2.-Data-Analysis
- https://www.researchgate.net/publication/350078499_Analyzing_Evolution_of_the_Olympics_by_Exploratory_Data_Analysis_using_R
- https://triemann.ca/wp-content/uploads/2021/01/Olympic-Analysis_Riemann_Nicol.pdf
- Data Analytics on Olympics Datasets (ijrpr.com)
- https://www.slideshare.net/irjetjournal/olympics-sports-data-analysis

APPENDIX - I

SOURCE CODE

GITHUB REPO:



GITHUB LINK: [HTTPS://GITHUB.COM/HITHACHOUDHARY/FUTURENSE-INTERNSHIP-CAPSTONE-PROJECT](https://github.com/hithachoudhary/Futurense-Internship-Capstone-Project)

SOURCE CODE SNIPPET:

Data Loading and Exploration

This step includes:
1.Reviewing the first lines of the data
2.Gathering datatypes,column names,statistical information and other information using the .describe() and .info() functions
3.Identifying missing values

Note: This dataset contains two csv files and hence we'll be exploring each file on it's own before merging it

In [2]:
```
athlete=pd.read_csv("athlete_events.csv")
noc=pd.read_csv("noc_regions.csv")
```

1.Athlete Events data

In [3]:
```
#Reviewing the dataframe

athlete_df=pd.DataFrame(athlete)
athlete_df
```

Out[3]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN
2	3	Gunnar Nielsen	M	24.0	NaN	NaN	Denmark	DEN	1920	1920	Summer	Antwerpen	Football	Football Men's	NaN

APPENDIX-II

DATASHEETS

"THE OLYMPICS" DATASET CONSISTS OF TWO CSV FILES:

ATHLETES.CSV:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
2	1	A Dijiang	M	24	180	80	China	CHN	1992 Sumr	1992	Summer	Barcelona	Basketball	Basketball	NA
3	2	A Lamusi	M	23	170	60	China	CHN	2012 Sumr	2012	Summer	London	Judo	Judo Men's	NA
4	3	Gunnar Nielsen Aab	M	24	NA	NA	Denmark	DEN	1920 Sumr	1920	Summer	Antwerpen	Football	Football M	NA
5	4	Edgar Lindenau Aab	M	34	NA	NA	Denmark/S	DEN	1900 Sumr	1900	Summer	Paris	Tug-Of-Wa	Tug-Of-Wa	Gold
6	5	Christine Jacoba Aa	F	21	185	82	Netherlanc	NED	1988 Winte	1988	Winter	Calgary	Speed Skat	Speed Skat	NA
7	5	Christine Jacoba Aa	F	21	185	82	Netherlanc	NED	1988 Winte	1988	Winter	Calgary	Speed Skat	Speed Skat	NA
8	5	Christine Jacoba Aa	F	25	185	82	Netherlanc	NED	1992 Winte	1992	Winter	Albertville	Speed Skat	Speed Skat	NA
9	5	Christine Jacoba Aa	F	25	185	82	Netherlanc	NED	1992 Winte	1992	Winter	Albertville	Speed Skat	Speed Skat	NA
10	5	Christine Jacoba Aa	F	27	185	82	Netherlanc	NED	1994 Winte	1994	Winter	Lillehamm	Speed Skat	Speed Skat	NA
11	5	Christine Jacoba Aa	F	27	185	82	Netherlanc	NED	1994 Winte	1994	Winter	Lillehamm	Speed Skat	Speed Skat	NA
12	6	Per Knut Aaland	M	31	188	75	United Stat	USA	1992 Winte	1992	Winter	Albertville	Cross Cour	Cross Cour	NA
13	6	Per Knut Aaland	M	31	188	75	United Stat	USA	1992 Winte	1992	Winter	Albertville	Cross Cour	Cross Cour	NA
14	6	Per Knut Aaland	M	31	188	75	United Stat	USA	1992 Winte	1992	Winter	Albertville	Cross Cour	Cross Cour	NA
15	6	Per Knut Aaland	M	31	188	75	United Stat	USA	1992 Winte	1992	Winter	Albertville	Cross Cour	Cross Cour	NA

THIS DATASET CONSISTS OF A TOTAL OF 271116 ROWS AND 17 COLUMNS

NOC_REGIONS.CSV

	A	B	C	D
1	NOC	region	notes	
2	AFG	Afghanistan		
3	AHO	Curacao	Netherlands Antilles	
4	ALB	Albania		
5	ALG	Algeria		
6	AND	Andorra		
7	ANG	Angola		
8	ANT	Antigua	Antigua and Barbuda	
9	ANZ	Australia	Australasia	

THIS DATASET CONSISTS OF A TOTAL OF 230 ROWS AND 3 COLUMNS